Classification, Analysis, and Prediction of the Daily Operations of Airports Using Machine Learning

Eugene Mangortey*, Tejas G. Puranik†, Olivia J. Pinon‡, and Dimitri N. Mavris§
Georgia Institute of Technology, Atlanta, GA, 30332

The Federal Aviation Administration (FAA) is the regulatory body in the United States responsible for the advancement, safety, and regulation of civil aviation. The FAA also oversees the development of the air traffic control system in the U.S. Over the years, the FAA has made tremendous progress in modernizing the National Airspace System (NAS) by way of technological advancements and the introduction of procedures and policies that have maintained the safety of the United States airspace. However, as with any other system, there is a need to continuously address evolving challenges pertaining to the sustainment and resiliency of the NAS. One of these challenges involves efficiently analyzing and assessing the operations of airports. In particular, there is a need to assess the impact and effectiveness of the implementation of Traffic Management Initiatives (TMI) and other procedures on daily airport operations, as this will lead to the identification of trends and patterns to inform better decision making. The FAA currently manually classifies the daily operations of airports into three categories: "Good Days", "Average Days", and "Bad Days" as a means to assess their efficiency. However, this exercise is time-consuming and can be improved. In particular, Big Data Analytics can be leveraged to develop a systematic approach for classifying or clustering the daily operations of airports. This research presents a methodology for clustering the daily operations of Newark International Airport (EWR) using metrics such as the number of diversions, Ground Stops, departure delays, etc. Each of these categories/clusters is then analyzed to identify key characteristics, trends and patterns, which can then be used by airport operators, and FAA analysts and researchers to improve the operations at the airport. Finally, the Boosting Ensemble Machine Learning algorithm is used to predict the category of operations at the airport, hence enabling airport operators, FAA analysts and researchers to take appropriate actions.

I. Nomenclature

AD	=	Average Distance
ADM	=	Average Distance between Means
APN	=	Average Proportion of Non-overlap
$ASPM$	=	Aviation System Performance Metrics
ATC	=	Air Traffic Controller
EWR	=	Newark International Airport
FAA	=	Federal Aviation Administration
FN	=	False Negative
FOM	=	Figure Of Merit
FP	=	False Positive
GDP	=	Ground Delay Program
GS	=	Ground Stop
NAS	=	National Airspace System
$OSPC$	=	Operational Service Performance Criteria

*Graduate Research Assistant, Aerospace Systems Design Laboratory, Daniel Guggenheim School of Aerospace Engineering, AIAA Student Member

†Research Engineer II, Aerospace Systems Design Laboratory, Daniel Guggenheim School of Aerospace Engineering, AIAA Member

‡Senior Research Engineer, Aerospace Systems Design Laboratory, Daniel Guggenheim School of Aerospace Engineering, AIAA Senior Member

§S.P. Langley NIA Distinguished Regents Professor and Director of Aerospace Systems Design Laboratory, Daniel Guggenheim School of Aerospace Engineering, AIAA Fellow

PCA = Principal Component Analysis
TMI = Traffic Management Initiatives
TN = True Negative
TP = True Positive
VAT = Visual Assessment of clustering Tendency

II. Introduction

The National Airspace System (NAS) refers to a network of airports, services, rules, regulations, and procedures used by the Federal Aviation Administration (FAA) to regulate and maintain air transportation in the United States [1]. Over the last decade, the FAA has successfully initiated and implemented initiatives aimed at modernizing the National Airspace System and improving aviation safety in the United States. These initiatives, primarily through the FAA's NextGen initiative, have involved the implementation of new technologies and procedures to ensure that the National Airspace System remains safe, efficient and resilient [2]. However, as with any other system, the NAS continually faces challenges that need to be addressed to ensure the safety of air transportation in the United States. In particular, a key challenge involves assessing the efficiency of airport operations in order to assess the impact and effectiveness of actions taken by traffic management personnel, which could in turn inform about trends and patterns. In this context, efforts have been pursued by analysts at the FAA to assess the operations of eight airports in the United States using the Operational Service Performance Criteria.

A. Operational Service Performance Criteria

The Operational Service Performance Criteria is used by analysts and researchers at the FAA to classify the daily operations of eight U.S. airports into three categories: "Good days", "Average days", and "Bad days". This classification is currently performed using the following parameters:

- **Traffic Management Initiatives (TMI) To Delays**: These are delays to airports caused by the implementation of Traffic Management Initiatives [3]. This parameter should be minimized
- **Departure delays**: Departure delays in excess of 15 minutes attributed to conditions at the departure airport [3]. This parameter should be minimized
- **GDP Revisions**: Ground Delay Programs are Traffic Management Initiatives implemented when aircraft demand is projected to exceed airport capacity over a long period of time [4–6]. This parameter refers to the number of times that the Ground Delay Program was updated. This parameter should be minimized
- **GDP Lead-in Time (Minutes)**: The time between the proposal of a Ground Delay Program and its implementation. This parameter should be maximized
- **Ground Stops**: These are Traffic Management Initiatives implemented when aircraft demand is projected to exceed airport capacity over a short period of time [6, 7]. This parameter refers to the number of Ground Stops implemented at the airport and should be minimized
- **Number of aircraft affected by airborne holding events**: Airborne holding occurs when an en-route aircraft is issued a clearance in excess of 15 minutes for a predetermined maneuver to keep the aircraft within a specified airspace while awaiting further clearance from Air Traffic Controllers. This parameter should be minimized [8]
- **Total duration of airborne holding events**: The summation of durations of all airborne holding events over 15 minutes. This parameter should be minimized
- **Diversions**: The number of flights that were diverted from their originally intended arrival airport. This parameter should be minimized
- **Completion rate**: The percentage of scheduled and/or planned air carrier arrivals that were not cancelled [8]. This parameter should be maximized

Each of these parameters is classified as green (good), yellow (average), or red (bad) using predefined ranges of values, as seen in Figure 1. The classification of the daily operation of an airport is then determined by identifying the predominant class of parameters (green, yellow, red) for the airport, as seen in Figure 1, where EWR, for example, was classified as a "Good day" because green (good) was the predominant class of parameters.

August XX, 20XX	Green	Yellow	Red	BOS	EWR	LGA	JFK	PHL	IAD	BWI	DCA
TMI To (including GS, GDPs, Other)	0-75	76-200	+201	0	211	260	88	134	0	0	60
Departure Delays	0-25	26-75	+75	0	90	262	60	152	36	30	22
GDP Revisions (*Manual Entry from NE Recap*)	0-1	2-3	+4	0	2	1	1	0	0	0	0
GDP Lead-in Time (Minutes)	+120	45-119	-45	n/a	146	0	143	105	n/a	n/a	0
Ground Stops (*Manual Entry from NE Recap*)	0-1	2-4	+5	0	0	1	1	1	0	0	1
Airborne Holding (Minutes)	0-75	76-200	+201	0	0	33	0	279	0	114	595
Airborne Holding (# of aircraft)	0-7	8-20	+21	0	0	2	0	10	0	5	25
Diversions	0-4	5-10	+11	0	1	1	1	1	0	1	9
Completion Rate	+90	80-90	-80	98.91	96.94	97.89	93.54	97.78	98.76	98.11	96.76
				G	G	G	G	Y	G	G	G

Fig. 1 Operational Service Performance Criteria for a particular day

As with any other system, the Operational Service Performance Criteria has a number of limitations that need to be addressed. First, this classification process can be time consuming as analysts have to manually input data on a daily basis. There is thus a need to develop a systematic approach for classifying or clustering the daily operations of airports.

Second, there is also a need to determine if classifying the daily operations of airports into three categories is the best suited approach for assessing airport operations. Machine Learning algorithms can be leveraged to determine the optimal number of categories for classifying airport operations.

Finally, the current approach of using the predominant class (green, yellow, red) of parameters to categorize the daily operations of airports assumes that each parameter is weighted equally. However, the impact of each parameter may vary on an airport-by-airport basis.

B. Review of other related efforts

Efforts have been made over the years by researchers to provide insights into and improve the efficiency of airport operations using Machine Learning. First, Bieblich et al. [9] developed a methodology for generating generic flight schedules. This methodology involved categorizing airports using clustering algorithms and using clusters or categories as inputs for determining generic flight schedules. This was achieved by categorizing airports based on the nature of their operations (cargo hub, small regional airports, etc). Even though their research did not provide insights into the operations of airports, it highlighted the use of clustering algorithms to group airports into clusters. This can thus be extended to using clustering algorithms to group the operations of an airport in order to identify and assess trends and patterns that were not previously known.

Second, Zambochova [10] grouped 838 airports into clusters based on the number of handled passengers using monthly data from January 2000 to April 2014. It is worth noting that some underlying factors tend to be airport/region specific so clustering at a regional airport level rather than many airports could be more insightful.

Finally, Grabbe and Sridhar [11] used clustering algorithms to identify hours for which the probability of imposing a Ground Delay Program were similar at the Chicago O'Hare International Airport and Newark Liberty International Airport. Ground Delay Programs (GDP) are utilized by controllers to manage air traffic whenever the number of anticipated aircraft is projected to exceed an airport's acceptance rate over a long period of time [12]. An analysis of the clusters was also conducted to identify the underlying weather conditions in each of these clusters. While the scope of their work primarily focused on Ground Delay Programs, it can be extended beyond GDPs to daily airport operations.

C. Research Objectives

The review of prior research and that of the Operational Service Performance Criteria highlight a few limitations and/or gaps. First, airport operations are impacted by a variety of factors (weather, volume etc.). Thus, grouping airports from around the world into clusters may not be the most appropriate approach, as underlying causes of events at airports may be lost. There is thus a need to focus on the operations of individual airports instead of a group of airports. This research focuses on the operations at Newark International Airport (EWR).

Second, it is also important to note the impact that the time of year has on the daily operations of airports. For example, clustering data containing summer and winter airport operations may not be appropriate as varying weather conditions often have significant impacts on airport operations. This research focuses on operations at Newark International Airport during the months of September, October, and November 2018. This process can be replicated for other seasons.

Third, the review of prior research also revealed that a rigorous benchmarking of different clustering algorithms is lacking. Clustering algorithms perform differently depending on the type and amount of data, as well as the algorithms'

methodology. There is thus a need to perform a benchmarking exercise to rigorously determine the most appropriate clustering algorithm for this specific problem.

Fourth, as mentioned previously, the FAA's Operational Service Performance Criteria can be improved by developing a systematic approach for classifying or clustering the daily operations of airports. There is also a need to determine if classifying the daily operations of airports into three categories is the best suited approach for assessing airport operations. Machine Learning algorithms can be leveraged to determine the optimal number of categories for classifying airport operations.

Finally, after clustering the daily operations of airports, and identifying the characteristics of each cluster, there is also a need to develop a Machine Learning model to predict which category a daily operation will fall into. This will make it easier for analysts and researchers to identify causes of the inefficiencies in the daily operations of airports. Based on the research gaps identified, the focus of this research is thus four-fold:

1) Identify the optimal number of categories for classifying daily airport operations
2) Categorize the daily operations of airports
3) Identify and analyze the characteristics of the categories to potentially reveal trends and patterns
4) Predict the category that a daily operation will belong to

The first, second, and third objectives are achieved by benchmarking different clustering algorithms and assessing their performance using clustering validation methods. The final objective is achieved by leveraging the Boosting Ensemble Machine Learning technique, based on its performance in other research efforts [5, 13, 14]. The remainder of this paper provides an overview of clustering algorithms, and discusses the methodology as well as the results obtained from this research.

III. Overview of Clustering Algorithms

Clustering is an unsupervised Machine Learning technique that categorizes data into clusters, without any prior training or understanding of the data [15–17]. Typically, items in a cluster should share similar characteristics but should be different from items in other clusters. The clustering algorithms benchmarked for this research are the Hierarchical, Kmeans, Partitioning Around Medoids, Fuzzy Analysis, Sota, Clara, and Model algorithms.

A. Hierarchical Clustering Algorithms

Hierarchical clustering algorithms group similar objects in multidimensional spaces into categories by either repeatedly dividing one cluster into at least two clusters (divisive) or by assigning each object into a cluster, and then merging similar clusters by their proximity to each other (agglomerative) [18].

1. Divisive Hierarchical Clustering Algorithms

The divisive hierarchical clustering algorithms utilized for this research are:

- Divisive Analysis (DIANA): This algorithm initially places all objects into the same cluster. At each point in time, the algorithm then splits the largest available cluster into two smaller clusters until each cluster contains at least one object [19]
- Self Organizing Tree Algorithm (SOTA) Clustering Algorithm: This algorithms splits objects into clusters without specifying a predetermined number of clusters. The algorithm splits the objects into clusters by detecting patterns in the dataset without any human interaction [20]

2. Agglomerative Hierarchical Clustering Algorithms

The agglomerative hierarchical clustering algorithms utilized for this research are:

- Complete Linkage: The distance between two clusters is defined as the longest distance between two objects in each cluster [21]
- Average Linkage: The distance between two clusters is defined as the average distance between each object in one cluster to every object in the other cluster [21]
- Centroid Linkage: The distance between two clusters is defined as the distance between the centroids of the clusters

- Single Linkage: The distance between two clusters is defined as the shortest distance between two objects in each cluster [21]
- Ward: The distance between two clusters is defined as how much the sum of squares will increase when the clusters are merged [22]

B. Kmeans Clustering Algorithm

The Kmeans algorithm assigns objects to a predetermined number of clusters, where the differences between objects in each cluster are minimized, and the differences between objects in different clusters are maximized [15].

C. Partitioning Around Medoids (PAM) Clustering Algorithm

The Partitioning Around Medoids (PAM) or k-medoids clustering algorithm is similar to the Kmeans clustering algorithm. However, the PAM algorithm clusters objects into a predetermined number of clusters around medoids or centers [19].

D. Fuzzy Analysis (FANNY) Clustering Algorithm

This algorithm clusters observations in such a manner that each object can belong to more than one cluster. Each object is then assigned a membership coefficient which indicates how much an object belongs to the different clusters [19].

E. Clustering for Large Applications (CLARA) Clustering Algorithm

This algorithm works similarly to the Partitioning Around Medoids (PAM) or k-medoids clustering algorithm, where objects are clustered around centers or medoids. However, the CLARA algorithm only clusters a sample of the large dataset, and then assigns the remaining objects in the dataset to the clusters obtained from the sample [19].

F. Model-based Clustering Algorithm

This algorithm is a statistical model made up of a combination of Gaussian distributions that are used to fit the data, where each combination of Gaussian distributions represents a cluster [23].

IV. Methodology

EWR airport data from 79 days between the months of September and November, 2018 was extracted in csv format from the FAA's Aviation Systems Performance Metrics (ASPM) database [24] and used for this research. Seven of the nine metrics used for this research were readily available in the database. The GDP Revisions and GDP Lead-in Time (Minutes) metrics were calculated using data available in the database. Figure 2 provides an overview of the methodology used for this research. This iterative process can be repeated with additional airport data and across different airports.

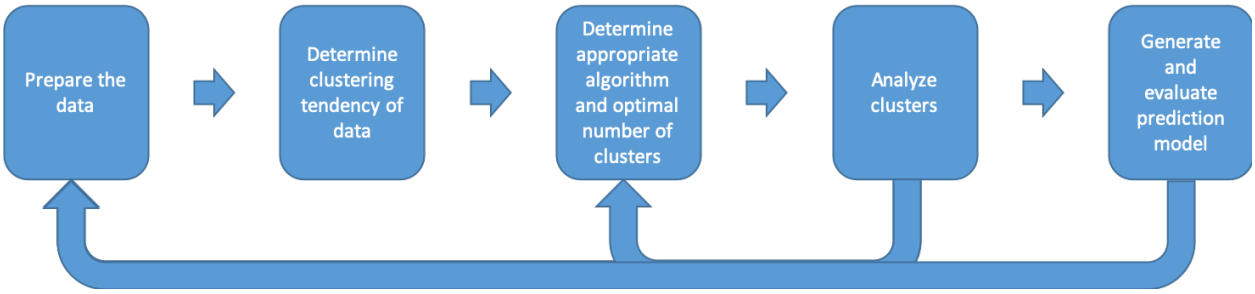


Fig. 2 Overview of methodology

This methodology was implemented using R [25] and is as follows:

A. Prepare the data

The following steps were taken to prepare the data for clustering:

1. Update GDP Lead-in Time (Minutes) parameter

GDP Lead-in Time is the time between the proposal of a Ground Delay Program, and its implementation. This variable is set as "n/a" whenever a Ground Delay Program is not implemented at an airport, as seen in Figure 1. There is thus a need to update "n/a" to a number as the majority of clustering algorithms do not allow the use of categorical data. As shown in Figure 1, higher GDP Lead-in times correspond to better operational performance. Consequently, "n/a" was updated to 500 minutes, based on feedback from Subject Matter Experts (SME) at the FAA, as a GDP will never be proposed 500 minutes prior to its implementation.

2. Z-score standardization

To ensure that parameters with larger ranges of values do not skew the clustering process, there is a need to normalize the parameters. Z-score standardization is used to scale parameters to ensure that they have a mean of zero and a standard deviation of one [26, 27]. This is achieved with the mean and standard deviation of the parameter, and is calculated using:

$$Z = \frac{Value - Mean}{Standard\ Deviation}$$

3. Principal Component Analysis (PCA)

It is usually difficult to explore and visualize the relationships between features in highly dimensional datasets. Thus, techniques such as Principal Component Analysis (PCA) have been widely used to reduce the dimensionality of datasets. This is achieved by orthogonally transforming a set of variables into principal components. Principal components are linear combinations of original variables of a dataset that capture the variance of the dataset. The transformation is done to ensure that the first principal component captures the maximum variance of the dataset, while each subsequent principal component captures the remaining variance of the dataset [28–31]. Orthogonal components that capture the maximum variance of the dataset are then identified using a scree plot, as seen in Figure 3. The scree plot, for this example, shows that 30 components captured about 98% of the variance of the dataset. This means that the dimensionality of the dataset is reduced from over 40 variables to 30 without significantly reducing the variance of the dataset.

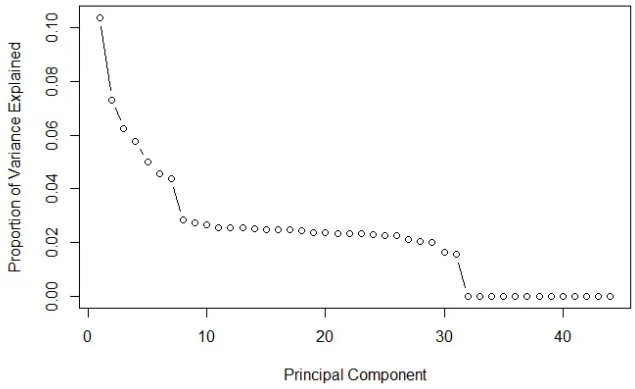


Fig. 3 Sample scree plot [32]

B. Determine the clustering tendency of the data

Majority of clustering algorithms split up datasets into predefined number of clusters, even if no meaningful clusters exist. It is thus important to assess the clustering tendency of a dataset to determine if meaningful clusters can be

created [33–36]. Hopkins statistic [37] and Visual Assessment of cluster Tendency (VAT) [33–36, 38] are two methods that are usually used to determine if a dataset has a non-random structure and will produce useful clusters [38]. The null hypothesis for the Hopkins statistic is defined as the dataset being uniformly distributed. The alternative hypothesis on the other hand is defined as the dataset not being uniformly distributed. Thus, a Hopkins statistic close to zero means that the null hypothesis is rejected and the dataset has a high clustering tendency [39]. The Visual Assessment of cluster Tendency (VAT) on the other hand is an image that indicates the presence of meaningful and well separated clusters, represented by dark boxes along the main diagonal of the image, as seen in Figure 4. The VAT is implemented by computing the Dissimilarity Matrix (DM) between objects in the dataset using the Euclidean distance measure [40, 41]. An Ordered Dissimilarity Matrix (ODM) is then created by reordering the original Dissimilarity Matrix so that similar objects are close to one other. The Ordered Dissimilarity Matrix is then displayed as the VAT for the dataset [33–36].

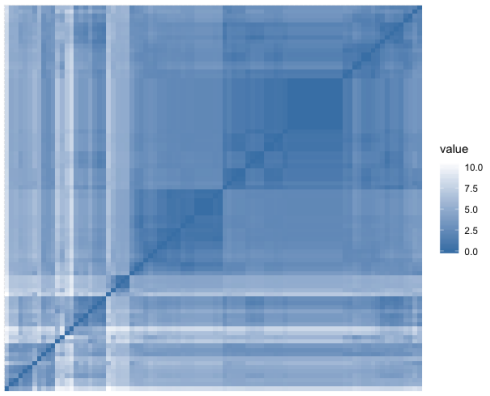


Fig. 4 Visual Assessment of cluster Tendency (VAT)

C. Determine the appropriate clustering algorithm and optimal number of clusters

Clustering algorithms have different methodologies and assumptions. It is thus important to benchmark different clustering algorithms in order to identify the best suited clustering algorithm for the dataset. This is achieved by using metrics that evaluate the consistency of clustering results [42, 43], assess cluster quality, and reflect the compactness, connectedness, and separation of the cluster partitions [20, 43, 44]. The following metrics are weighted equally for the purpose of this research:

- **Connectivity:** "This measures the extent to which items are placed in the same cluster as their nearest neighbors in the data space" [20, 43, 45]. Connectivity ranges from zero to infinity and should be minimized
- **Dunn Index:** This measures the ratio between the smallest distance between items in different clusters and the largest distance between items in the same cluster [43, 45, 46]. The Dunn Index ranges from zero to infinity and should be maximized
- **Silhouette:** This measures the average distance between different clusters [43, 45, 47] and ranges from 1 to -1. A Silhouette score of -1 refers to poorly clustered items while a score of 1 refers to well clustered items
- **Average Proportion of Non-overlap (APN):** "This measures the ratio of items placed in different clusters by clustering using the entire dataset and clustering using the dataset with one excluded column" [20, 43]. APN ranges from 0 to 1, and should be minimized
- **Average Distance (AD):** "This measures the average distance between items placed in the same cluster when the entire dataset is clustered, and when the dataset is clustered without one column" [20, 43]. AD ranges from 0 and infinity, and should be minimized
- **Average Distance between Means (ADM):** "This measures the average distance between cluster centers for items in the same cluster when the entire dataset is clustered, and when the dataset is clustered without one column" [20]. ADM ranges from 0 to 1, and should be minimized
- **Figures of Merit (FOM):** "This measures the average intra-cluster variance of the deleted column, where the clustering is based on the remaining (undeleted) columns" [20, 43]. FOM ranges from 0 to 1, and should be minimized

Some clustering algorithms also require users to arbitrarily select the number of clusters to be used. In order to ensure that the optimal number of clusters is selected, there is a need to vary the number of clusters used while benchmarking the different algorithms to identify a combination of appropriate algorithm and optimal number of clusters to be used for this research.

D. Analyze clusters

After identifying the appropriate algorithm and the optimal number of clusters, clusters can then be generated and analyzed to identify the parameters that best represent each cluster. This analysis enables analysts and researchers to identify trends and patterns in the daily operations of the airport.

E. Generate and evaluate prediction model

The final objective of this research involves predicting the cluster that the daily operation of an airport would fall into. This can be achieved by leveraging the Boosting Ensemble Machine Learning algorithm, because of its performance in other research efforts [5, 13, 14]. The Boosting Ensemble algorithm builds a prediction model with training data, and then builds subsequent models that relearn incorrect predictions of the previous models until the final model accurately predicts all of the training data [15, 48–52]. The predictors for the models are the metrics listed in Section II, and the target is the cluster that a daily operation belongs to.

The data is randomly divided into two sets: training and testing. Three-fourths of the data is assigned to the training set, which is used to generate the model. The remainder of the data is used to evaluate the performance of the model. The model is evaluated using results obtained from a confusion matrix. A confusion matrix, as seen in Table 1, is a table that categorizes predictions according to whether they match the actual value.

Table 1 Confusion Matrix

	Actual: No	**Actual: Yes**
Predicted: No	True Negative (TN)	False Negative (FN)
Predicted: Yes	False Positive (FP)	True Positive (TP)

True Positive (TP) refers to the correct classification of the class of interest. True Negative (TN) refers to the correct classification of the class that is not of interest. False Positive (FP) refers to the incorrect classification of the class of interest. False Negative (FN) refers to the incorrect classification of the class that is not of interest [15]. The performance of the models is finally analyzed by computing the Kappa Statistic, which accounts for the probability of a correct prediction by chance alone. Kappa Statistic is specified below where P_0 is the observed value and P_E is the expected value [53]:

$$K = \frac{P_0 - P_E}{1 - P_E}$$

Table 2 provides an interpretation of Kappa Statistic values [15].

Table 2 Interpretation of Kappa Statistic values

Kappa Statistic	**Interpretation**
< 0.2	Poor Agreement
0.2 - 0.4	Fair Agreement
0.4 - 0.6	Moderate Agreement
0.6 - 0.8	Good Agreement
0.8 - 1	Very Good Agreement

V. Results

A. Principal Component Analysis (PCA)

Principal Component Analysis was used to reduce the dimensionality of the dataset from 9 parameters to 4 principal components that captured about 90% of the variance of the dataset, as seen in Figure 5. These principal components were used for the remainder of this research.

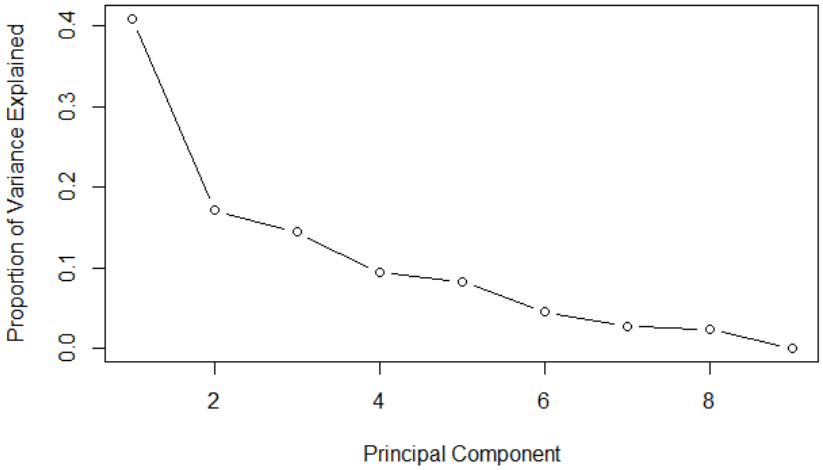


Fig. 5 Scree plot from Principal Component Analysis

B. Determine the clustering tendency of the data

The clustering tendency of the data was evaluated using Hopkins Statistic and the Visual Assessment of cluster Tendency (VAT). The Hopkins statistic obtained was **0.168**, which indicates high clustering tendency of the data. The presence of dark boxes along the diagonal in Figure 6 also indicates that the data is "clusterable".

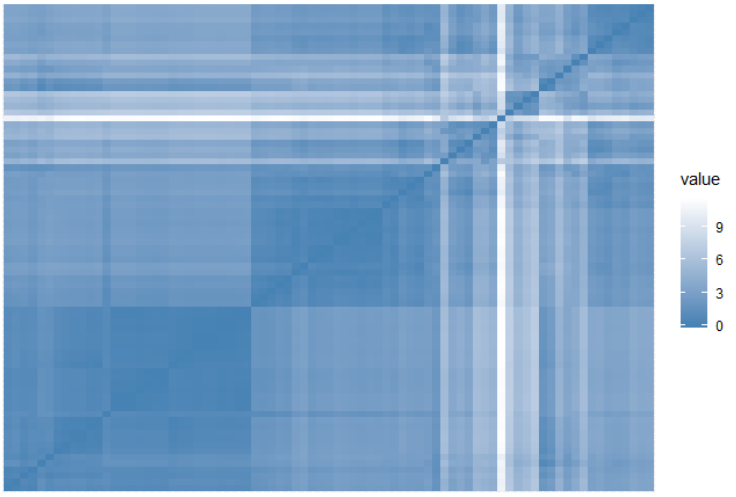


Fig. 6 Visual Assessment of cluster Tendency (VAT) for the data

C. Determine the appropriate clustering algorithm(s) and optimal number of clusters

Figure 7 shows the results obtained from the validation methods used to determine the appropriate clustering algorithm and optimal number of clusters. In particular, it shows that the most frequent optimal combination of algorithm and number of clusters is one of the Hierarchical algorithms, and two clusters. Consequently, this combination was identified to be the best suited for the data. It can also be seen that the SOTA algorithm was unable to create six clusters from the dataset. This can be explained by the fact that the SOTA algorithm splits objects by detecting patterns in the dataset. Thus, the algorithm did not detect patterns that could lead to the creation of six clusters.

```
Cluster sizes:
 2 3 4 5 6


Algorithm    Metric          2       3       4       5       6

hierarchical APN           0.0083  0.0887  0.0775  0.2762  0.4132
             AD            2.7289  2.4852  2.3144  2.2058  2.1659
             ADM           0.0651  0.5496  0.6173  0.9771  1.2657
             FOM           1.3034  1.2579  1.2193  1.1466  1.1059
             Connectivity  2.9290 12.4071 19.2147 22.7754 24.5587
             Dunn          0.8907  0.2466  0.2343  0.2343  0.2343
             Silhouette    0.7155  0.4947  0.4755  0.4266  0.3965
kmeans       APN           0.0236  0.1431  0.1799  0.2082  0.2302
             AD            2.4502  2.3015  1.8374  1.7410  1.5806
             ADM           0.5416  0.7265  0.7920  0.8415  0.8546
             FOM           1.3642  1.3024  1.1117  1.0396  1.0189
             Connectivity 13.8575 16.0714 12.8071 22.0052 29.8687
             Dunn          0.0609  0.0630  0.1677  0.1123  0.1657
             Silhouette    0.4449  0.4365  0.5030  0.4944  0.5182
pam          APN           0.2118  0.1217  0.1193  0.1561  0.1981
             AD            2.4167  1.8586  1.6269  1.5281  1.3885
             ADM           0.9211  0.5727  0.6325  0.6530  0.6349
             FOM           1.4018  1.2504  1.1089  1.0954  1.0160
             Connectivity  2.7480 14.6603 19.8266 27.6988 29.0810
             Dunn          0.0847  0.0729  0.0435  0.0907  0.1392
             Silhouette    0.4027  0.4443  0.4867  0.5025  0.5224
diana        APN           0.0551  0.1582  0.1565  0.1807  0.2231
             AD            2.7802  2.4377  1.9082  1.8144  1.6424
             ADM           0.3248  0.6691  0.8405  0.8530  0.8835
             FOM           1.3668  1.2325  1.0755  1.0206  1.0039
             Connectivity  2.9290 15.6433 16.0433 19.6040 26.2865
             Dunn          0.8907  0.2128  0.1588  0.1677  0.2290
             Silhouette    0.7155  0.4927  0.5003  0.4909  0.5110
sota         APN           0.2783  0.2869  0.2907  0.2785      NA
             AD            2.5491  2.2776  2.2007  2.0746      NA
             ADM           1.2133  1.1877  1.2956  1.2050      NA
             FOM           1.4257  1.3643  1.3418  1.2682      NA
             Connectivity 13.0456 28.3091 29.6976 31.6476      NA
             Dunn          0.0173  0.0184  0.0237  0.0374      NA
             Silhouette    0.3764  0.3965  0.4245  0.4266      NA
clara        APN           0.1104  0.1334  0.1755  0.2265  0.2814
             AD            2.2722  1.8886  1.6909  1.6055  1.4734
             ADM           0.5161  0.6178  0.7227  0.7911  0.8528
             FOM           1.3562  1.2528  1.2533  1.1001  0.9831
             Connectivity  2.7480 14.5421 22.3345 29.0123 26.7087
             Dunn          0.0847  0.0743  0.0306  0.0202  0.0904
             Silhouette    0.4027  0.4379  0.4659  0.3819  0.5128
model        APN           0.1077  0.2120  0.2745  0.3199  0.3157
             AD            2.3852  2.1104  2.0155  1.7453  1.5582
             ADM           0.3679  0.7991  1.0386  0.9642  0.8037
             FOM           1.2858  1.2313  1.1438  1.1465  1.0784
             Connectivity 14.7837 24.3274 33.9480 26.9778 40.2389
             Dunn          0.0282  0.0283  0.0283  0.0602  0.0386
             Silhouette    0.3213  0.3354  0.3782  0.4318  0.4048

Optimal Scores:

Metric       Score   Method       Clusters
APN          0.0083  hierarchical    2
AD           1.3885  pam             6
ADM          0.0651  hierarchical    2
FOM          0.9831  clara           6
Connectivity 2.7480  pam             2
Dunn         0.8907  hierarchical    2
Silhouette   0.7155  hierarchical    2
```

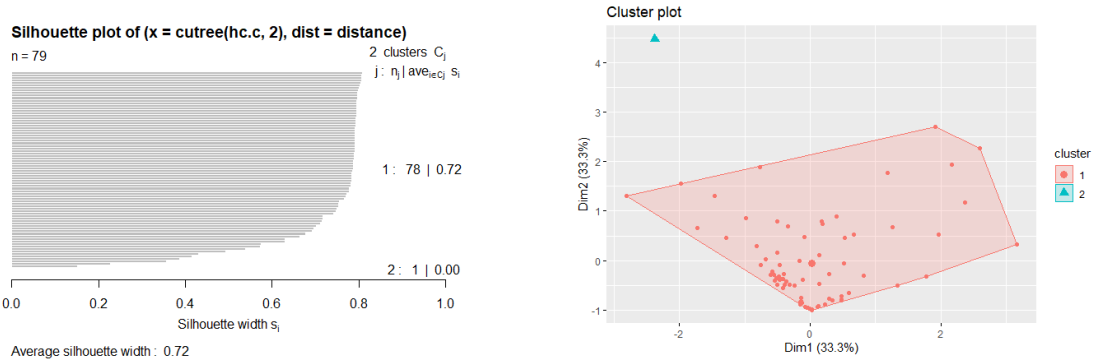
Fig. 7 Identification of appropriate clustering algorithm and optimal number of clusters

As mentioned previously, the Hierarchical group of algorithms is comprised of five individual algorithms: Complete Linkage, Single Linkage, Average Linkage, Centroid Linkage, and Ward algorithms. It was thus important to identify

the best suited Hierarchical algorithm for this research. This was achieved by developing clusters using these algorithms and analyzing their results using the Silhouette plot and width. A Silhouette plot shows how closely members in a cluster are connected to each other, with the maximum and minimum values of the Silhouette width (distance between members in a cluster) being 1 and -1, respectively, where 1 represents perfect clustering and -1 represents imperfect clustering.

The Average Linkage, Complete Linkage, Centroid Linkage and Single Linkage algorithms clustered daily operations of the Newark International Airport similarly. Figure 8a shows the Silhouette plot for these algorithms. It can be seen that 78 days were placed in one cluster while 1 day was an outlier. The cluster had a Silhouette score of **0.72**, which shows that the daily operations in this cluster were similar, compared to the outlier.

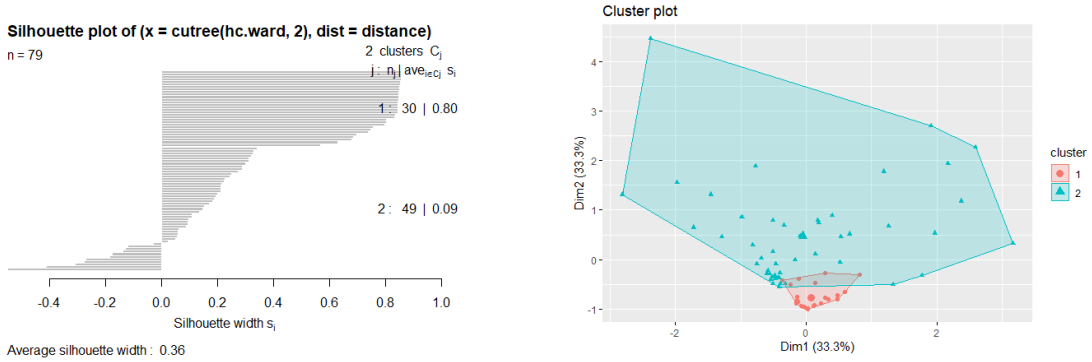
Figure 8b shows the main cluster and the outlier generated using the Average Linkage, Complete Linkage, Centroid Linkage, and Single Linkage clustering algorithms. The axes represent the first two principal components obtained from Principal Component Analysis, and were used as a means to visualize the clusters.



(a) Silhouette Plot **(b) Representation of clusters**

Fig. 8 Silhouette plot and representation of clusters obtained from the Average Linkage, Complete Linkage, Centroid Linkage and Single Linkage clustering algorithms

Figure 9a shows that the Ward algorithm placed 30 daily operations in one cluster, and placed 49 daily operations in a second cluster. The first cluster had a Silhouette score of 0.80, which shows that the days in this cluster were very similar. The second cluster had a poor Silhouette score of 0.09, which shows that the days in this cluster were very dissimilar. This algorithm produced an overall Silhouette score of 0.39. Figure 9b illustrates the two clusters generated using Principal Component Analysis.



(a) Silhouette Plot **(b) Representation of clusters**

Fig. 9 Silhouette plot and representation of clusters obtained from the Ward clustering algorithm

Based on Figures 8 and 9, it can be concluded that the best suited algorithms for clustering daily operations of the Newark International Airport from September to November, 2018 were identified to be the **Average Linkage, Complete Linkage, Centroid Linkage and Single Linkage clustering algorithms**, since they had the highest Silhouette scores (0.72). However, these algorithms clustered 78 days into one cluster and produced an outlier. Analysis of this outlier, as

seen in Figure 10, revealed that Newark Liberty International Airport was severely constrained on that day. In particular, it shows that the outlier had a higher number of departure delays and delays caused by the implementation of Traffic Management Initiatives at the airport, compared to the average number of departure delays and, delays caused by the implementation of Traffic Management Initiatives (TMI) of all of the other days in the main cluster.

Cluster	Number of delays due to TMI at airport	Number of departure delays	Number of GDP Revisions	GDP lead-in times (minutes)	Number of Ground Stops	Duration of airborne delays (minutes)	Number of aircraft affected by airborne delays	Number of flight diversions	Completion rate (%)
1	106.4	26.8	0.24	135.1	0.333	94.9	4.72	1.05	98.3
2	111	75	3	143	3	367	19	16	68.1

Fig. 10 Mean values of variables in clusters

In order to further analyze the data, there was a need to either remove the outlier or use more than two clusters. The Hierarchical group of algorithms and 3 clusters were identified as the next best combination of appropriate algorithm and optimal number of clusters by the majority of validation methods, as seen in Figure 7. Figure 11 shows the 3 clusters generated by the Hierarchical group of algorithms. It also shows that the daily operation identified as an outlier from the previous exercise with two clusters was identified as an outlier by four out of five algorithms. It also shows that the Average and Centroid Linkage algorithms clustered the daily operation of Newark International Airport similarly. The Complete Linkage, Single Linkage and Ward algorithms had Silhouette scores of 0.45, 0.39, and 0.42, respectively. On the other hand, the Average and Centroid Linkage algorithms had Silhouette scores of **0.49**. Thus, the best suited algorithms for clustering the daily operations of Newark International Airport from September to November, 2018 are the **Average and Centroid Linkage algorithms**.

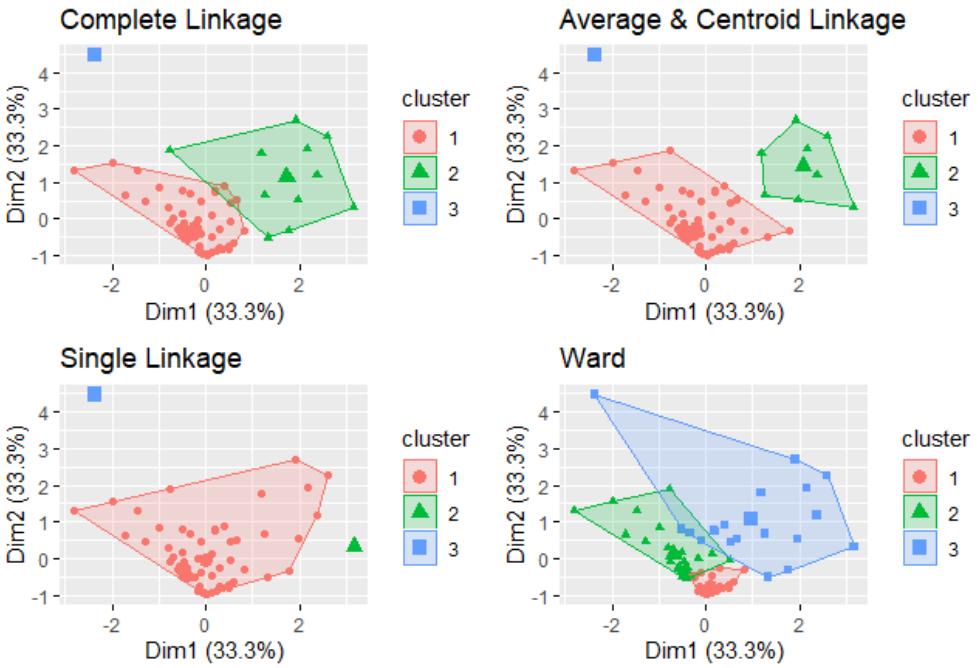


Fig. 11 Representation of clusters from the Hierarchical group of algorithms

D. Analyze clusters

Figure 11 shows that the Average and Centroid Linkage algorithms split the daily operations of EWR into two clusters and an outlier. The first and second clusters were comprised of 70 and 8 daily operations, respectively. Figure 12 shows the distribution of parameters across both clusters and the outlier. The outlier is characterized by high airborne holdings (minutes and number of aircraft), diversions, departure delays, GDP revisions, GDP lead-in time, and Ground Stops. It is also characterized by low completion rate and a moderate number of delays caused by the implementation of Traffic Management Initiatives to the airport. These characteristics correspond to poor operational performance.

The first cluster is generally characterized by low airborne holdings (minutes and number of aircraft), diversions, departure delays, GDP revisions, and Ground Stops. It is also characterized by high completion rate, and a wide range of GDP lead-in times and delays caused by the implementation of Traffic Management Initiatives at the airport. Overall, these characteristics correspond to good operational performance.

The second cluster is generally characterized by low diversions and Ground Stops. It is also characterized by a wide range of delays caused by the implementation of Traffic Management Initiatives at the airport, departure delays and GDP revisions, and high completion rates, GDP lead-in times and airborne holdings (minutes and number of aircraft). Overall, these characteristics correspond to varying operational performance where one or more parameters indicated sub-optimal to poor operational performance on the day.

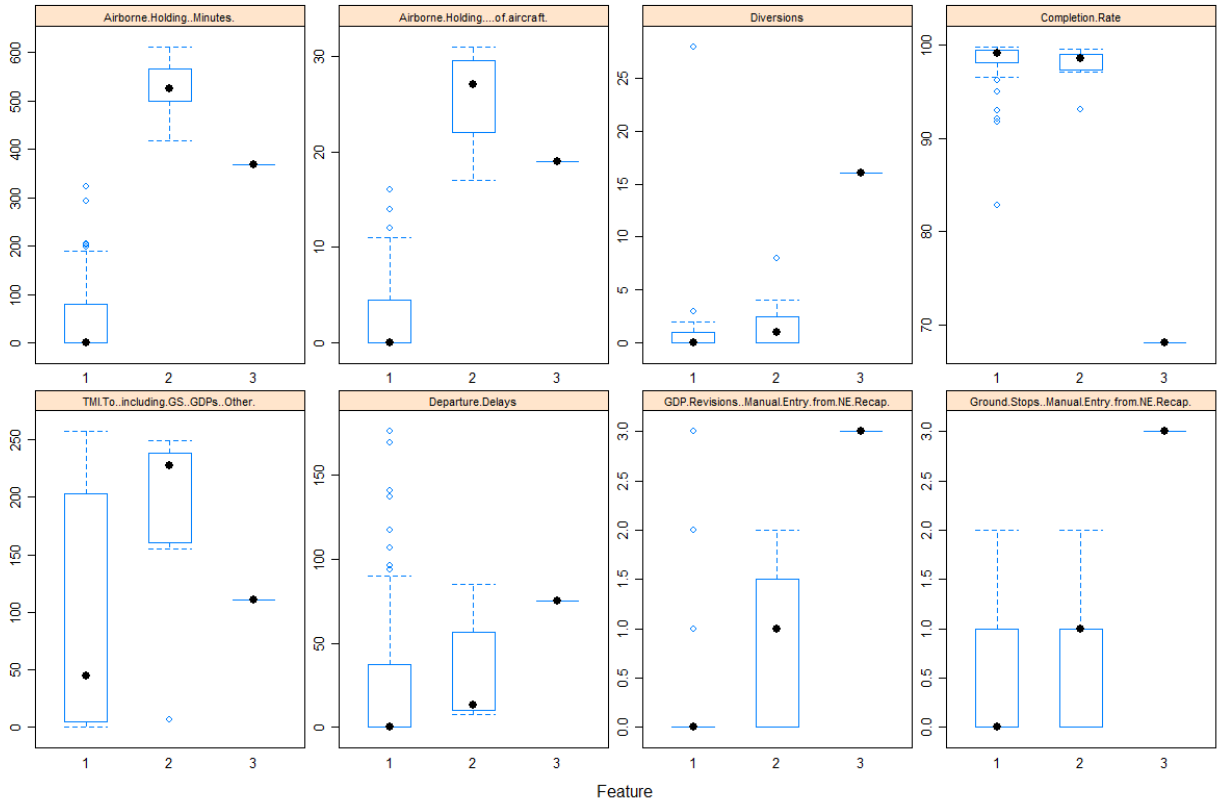


Fig. 12 Box plot showing the distribution of parameters across clusters

E. Comparison of Clustering approach and Operational Service Performance Criteria

This section provides a comparison of the classification of the daily operations of Newark Liberty International Airport using clustering and the FAA's Operational Service Performance Criteria (OSPC). Figure 13 shows that 69 days classified as "Good days" by OSPC were placed in the cluster characterized by good operational performance, while 2 days classified as "Average days" by OSPC were placed in the cluster characterized by sub-optimal to poor operational performance.

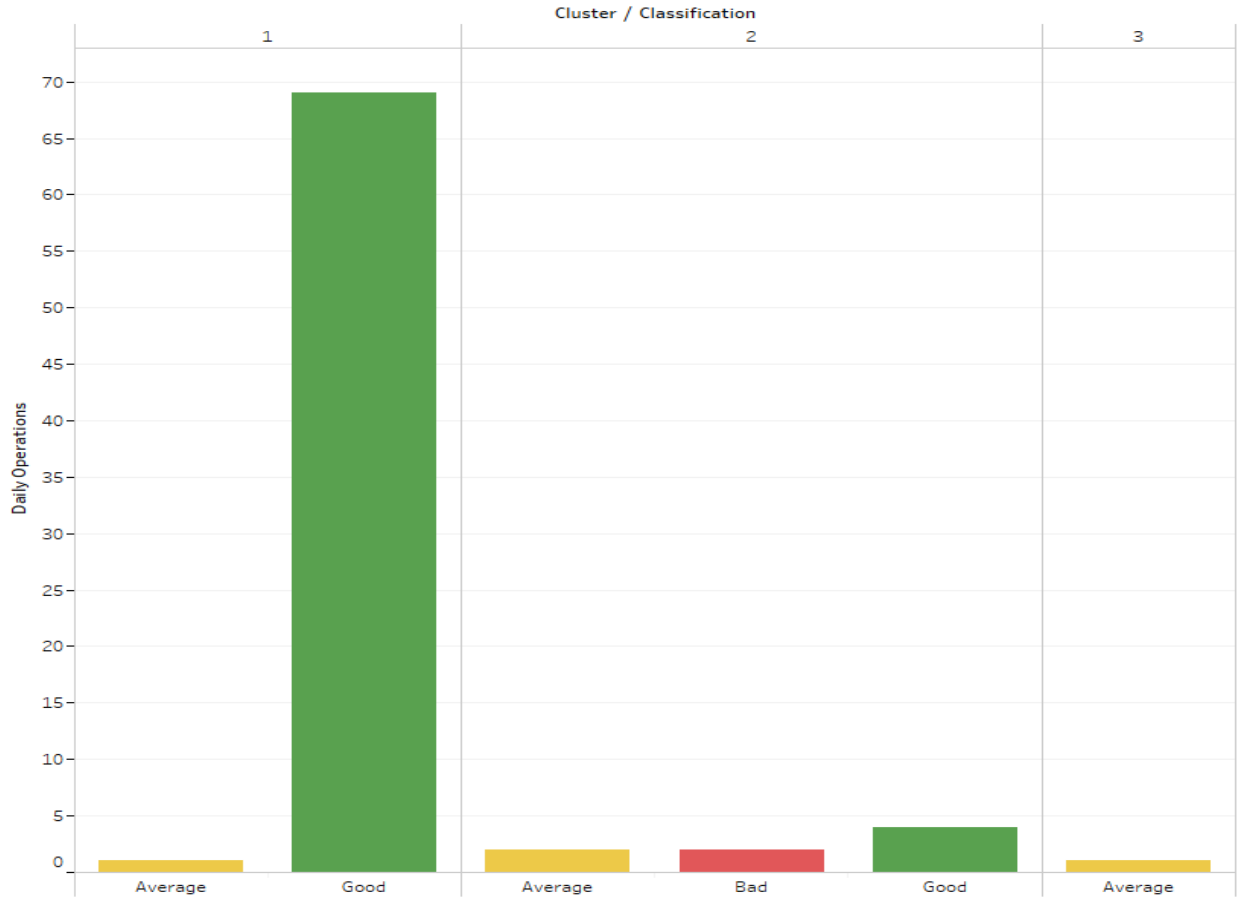


Fig. 13 Comparison of results from clustering and Operational Service Performance Criteria

The remaining 8 days were classified differently by both methods, as seen in Figure 14. In particular, it shows that days 2, 3, 4, and 8 were classified as "Good days" by OSPC. However, clustering classified these days into the second cluster, as they were characterized by high airborne holdings which severely impacted airport operations. OSPC also classified day 6 as an "Average Day" even though it was characterized by a very poor completion rate, as 32% of scheduled and/or planned air carrier arrivals were cancelled, due to severe constraints at the airport on that day. Days 1, 5, and 7 were classified differently by both methods due to varying operational performance where one or more parameters influenced their classification.

Day	Cluster	OSPC Classification	Number of delays due to TMI at airport	Number of departure delays	Number of GDP Revisions	GDP lead-in times (minutes)	Number of Ground Stops	Duration of airborne delays (minutes)	Number of aircraft affected by airborne delays	Number of flight diversions	Completion rate (%)
1	2	Bad	227	13	1	180	2	612	31	8	93.07
2	2	Good	155	11	0	162	0	417	19	0	98.76
3	2	Good	7	8	0	n/a	0	509	28	1	99.21
4	2	Good	249	31	2	184	1	603	31	0	97.68
5	1	Average	104	117	0	0	1	204	9	1	99.69
6	3	Average	111	75	3	143	3	367	19	16	68.06
7	2	Bad	236	82	0	87	0	525	25	0	99.52
8	2	Good	27	137	0	n/a	1	322	16	2	98.9

Fig. 14 Comparison of results from clustering and Operational Service Performance Criteria

Based on this analysis, it can be seen that using predefined ranges of values to classify the daily operations of airports is not the best approach for the task at hand as analysts would continually have to update these ranges based on prior knowledge and experience, instead of using a systematic approach such as clustering. However, this research validates the FAA's current classification of the daily operations of Newark Liberty International Airport into three categories.

F. Develop and evaluate prediction models

The final part of this research involves predicting the category that a daily operation belongs to using the Boosting Ensemble Machine Learning algorithm. This is achieved by randomly splitting the data into two sets. The training and testing sets contain 59 and 20 data points, respectively.

Table 3 shows the confusion matrix obtained from the Boosting Ensemble algorithm. The algorithm accurately predicted 17 days in Cluster 1, and 2 days in Cluster 2. However, it inaccurately predicted 1 day in Cluster 2 instead of Cluster 1. The model had a Kappa Statistic score of 0.773. It is important to note that the performance of the algorithm is limited by the amount of data. In this case, the algorithm's performance will be improved by obtaining and using additional airport data.

Table 3 Confusion matrix

	Actual Cluster 1	**Actual Cluster 2**	**Actual Outlier**
Predicted Cluster 1	17	0	0
Predicted Cluster 2	1	2	0
Predicted Outlier	0	0	0

VI. Conclusion & Future Work

This research proposes a systematic approach for categorizing the daily operations of airports using clustering algorithms, with a focus on Newark International Airport. This was done using metrics such as number of delayed flights, number of airborne-holding events, etc. The methodology involves assessing the clustering tendency of the data, identifying the most appropriate algorithm and optimal number of clusters, and analyzing the developed clusters to identify metrics that characterize or describe each cluster. The final step of the methodology involves using the Boosting Ensemble Machine Learning algorithm to predict the cluster that the daily operation of the airport will belong to. This methodology showed that the Average Linkage and Centroid Linkage clustering algorithms are the best suited algorithms for the data, while the optimal number of clusters was identified to be two. However, three clusters were used in order to gain further insights into the clusters. The clusters were then analyzed to identify their respective

characteristics. Finally, the Boosting Ensemble Machine Learning algorithm was used to predict the cluster that the daily operation of the airport belongs to.

Future work will focus on identifying thresholds of parameter values for each cluster and determining if the classification of airport operations should be on a seasonal basis. It will also focus on identifying the best suited algorithm and optimal number of clusters for the seven other airports currently assessed by the Operational Service Performance Criteria. Analysis of the results from this exercise will then be used to determine if the classification of the daily operations of airports should be on an airport level or on an airport-type level. Finally, future work will also involve fusing other datasets such as weather and air traffic data in order to identify trends and patterns, and the underlying factors leading to the characteristics of the daily operations of airports.

Acknowledgments

This research was carried out as a result of the collaboration between the Aerospace Systems Design Laboratory (ASDL) at Georgia Tech and the Federal Aviation Administration's Technical Center. Special thanks goes to Anya Berges and Warren Strickland from the Federal Aviation Administration's Technical Center for providing valuable feedback. The views and findings expressed in this document are those of the authors only, and do not represent those of the FAA.

References

[1] Dillingham, G., "National Airspace System: FAA Has Implemented Some Free Flight Initiatives, but Challenges Remain," *General Accounting Office, Washington DC*, 1998. URL https://www.gao.gov/assets/160/156365.pdf.

[2] Federal Aviation Administration, "What is NextGen?" , 2018. URL https://www.faa.gov/nextgen/what_is_nextgen/.

[3] Federal Aviation Administration, "OPSNET Reports: Definitions of Variables," , 2018. URL https://aspmhelp.faa.gov/index.php/OPSNET_Reports:_Definitions_of_Variables.

[4] Mangortey, E., Gilleron, J., Dard, G., Pinon-Fischer, O., and Mavris, D., "Development of a Data Fusion Framework to support the Analysis of Aviation Big Data," *AIAA Scitech 2019 Forum, AIAA SciTech Forum, (AIAA 2019-1538)*, 2019. URL https://doi.org/10.2514/6.2019-1538.

[5] Mangortey, E., Pinon, O., Puranik, T., and Mavris, D., "Predicting The Occurrence of Weather And Volume Related Ground Delay Programs," *AIAA AVIATION Forum*, 2019. URL https://doi.org/10.2514/6.2019-3188.

[6] Mangortey, E., Bleu Laine, M., Puranik, T., Pinon, O., and Mavris, D., "Application of Machine Learning to the Analysis and Prediction of the Coincidence of Ground Delay Programs and Ground Stops," *AIAA Science and Technology Forum (AIAA Scitech)*, 2020.

[7] Banavar, S., Gano, C., Shon, G., and Kapil, S., "Integration of Traffic Flow Management Decisions," *AIAA Guidance, Navigation, and Control Conference and Exhibit, Guidance, Navigation, and Control and Co-located Conferences*, 2002. URL https://doi.org/10.2514/6.2002-5014.

[8] Federal Aviation Administration, "ASPM Efficiency: Definitions of Variables," , 2018. URL https://aspmhelp.faa.gov/index.php/ASPM_Efficiency:_Definitions_of_Variables.

[9] Bießlich, P., Reitmann, S., Gollnick, V., Lütjens, K., Nachtigall, K., and Marx, S., "Developing Generic Flight Schedules for Airport Clusters," *5th CEAS Air and Space Conference*, 2015. URL https://elib.dlr.de/95309/.

[10] Marta, Z., "Cluster analysis of world's airports on the basis of number of passengers handled (case study examining the impact of significant events)," *Statistika*, Vol. 97, 2017, pp. 74–88. URL https://www.czso.cz/documents/10180/45606527/32019717q1074.pdf/3e368588-3138-4b8f-b977-184a4571b95e?version=1.0.

[11] Grabbe, S., and Sridhar, B., "Clustering Days with Similar Airport Weather Conditions," *14th AIAA Aviation Technology, Integration, and Operations Conference*, 2014. URL https://doi.org/10.2514/1.I010212.

[12] Ball, M., and Guglielmo, L., "Ground Delay Programs: Optimizing over the Included Flight Set Based on Distance," *14th AIAA Aviation Technology, Integration, and Operations Conference*, 2014. URL https://doi.org/10.2514/atcq.12.1.1.

[13] Dard, G., Mangortey, E., Pinon, O., and Mavris, D., "Application Of Data Fusion And Machine Learning To The Analysis Of The Relevance Of Recommended Flight Reroutes," *AIAA AVIATION Forum*, 2019. URL https://doi.org/10.2514/6.2019-3189.

[14] Mangortey, E., Puranik, T., Pinon, O., and Mavris, D., "Prediction and Analysis of Ground Stops with Machine Learning," *AIAA Science and Technology Forum (AIAA Scitech)*, 2020.

[15] Lantz, Brett, *Machine Learning with R: Discover How to Build Machine Learning Algorithms, Prepare Data, and Dig Deep into Data Prediction Techniques with R*, Packt Publishing, 2015. URL https://books.google.com/books?id=ZaJNCgAAQBAJ.

[16] Steinbach, M., Karypis, G., Kumar, V., et al., "A comparison of document clustering techniques," *KDD workshop on text mining*, Vol. 400, Boston, 2000, pp. 525–526.

[17] Johnson, S. C., "Hierarchical clustering schemes," *Psychometrika*, Vol. 32, No. 3, 1967, pp. 241–254. URL https://doi.org/10.1007/BF02289588.

[18] Salvador, S., and Chan, P., "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms," *6th IEEE International Conference on Tools with Artificial Intelligence,pp. 576-584,doi: 10.1109/ICTAI.2004.50*, 2004.

[19] Struyf, A., Hubert, M., and Rousseeuw, P., "Clustering in an Object-Oriented Environment," *Journal of Statistical Software, Articles*, Vol. 1, No. 4, 1997, pp. 1–30. doi:10.18637/jss.v001.i04, URL https://www.jstatsoft.org/v001/i04.

[20] Brock, G., Pihur, V., Datta, S., and Datta, S., "clValid: An R Package for Cluster Validation," *Journal of Statistical Software, Articles*, Vol. 25, No. 4, 2008, pp. 1–22. doi:10.18637/jss.v025.i04, URL https://www.jstatsoft.org/v025/i04.

[21] Olson, C., "Parallel algorithms for hierarchical clustering," *Parallel Computing 21, pp. 1313-1325*, 1995. URL https://doi.org/10.1016/0167-8191(95)00017-I.

[22] Murtagh, F., and Legendre, P., "Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm," *arXiv e-prints*, 2011. URL https://doi.org/10.1007/s00357-014-9161-z.

[23] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L., "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, Vol. 17, No. 10, 2001, pp. 977–987. URL https://doi.org/10.1093/bioinformatics/17.10.977.

[24] Analytics Vidhya, "FAA Operations & Performance Data," , 2019. URL https://aspm.faa.gov/.

[25] Ihaka, R., and Gentleman, R., "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics, 5:3, 299-314, DOI: 10.1080/10618600.1996.10474713*, 1996.

[26] Jain, A., Nandakumar, K., and Ross, A., "Score normalization in multimodal biometric systems," *Pattern Recognition*, Vol. 38, No. 12, 2005, pp. 2270 – 2285. doi:https://doi.org/10.1016/j.patcog.2005.01.012, URL http://www.sciencedirect.com/science/article/pii/S0031320305000592.

[27] Patro, S., and Sahu, K. K., "Normalization: A preprocessing stage," *arXiv preprint arXiv:1503.06462*, 2015.

[28] Jolliffe, I., *Principal component analysis*, Springer, 2011. URL https://link.springer.com/book/10.1007%2Fb98835.

[29] Wold, S., Esbensen, K., and Geladi, P., "Principal component analysis," *Chemometrics and intelligent laboratory systems*, Vol. 2, No. 1-3, 1987, pp. 37–52. URL https://doi.org/10.1016/0169-7439(87)80084-9.

[30] Abdi, H., and Williams, L. J., "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, Vol. 2, No. 4, 2010, pp. 433–459. URL https://doi.org/10.1002/wics.101.

[31] Candès, E. J., Li, X., Ma, Y., and Wright, J., "Robust Principal Component Analysis?" *J. ACM*, Vol. 58, No. 3, 2011, pp. 11:1–11:37. doi:10.1145/1970392.1970395, URL http://doi.acm.org/10.1145/1970392.1970395.

[32] Analytics Vidhya, "Practical Guide to Principal Component Analysis (PCA) in R & Python," , 2016. URL https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/.

[33] Bezdek, J. C., Hathaway, R. J., and Huband, J. M., "Visual assessment of clustering tendency for rectangular dissimilarity matrices," *IEEE Transactions on fuzzy systems*, Vol. 15, No. 5, 2007, pp. 890–903.

[34] Huband, J. M., Bezdek, J. C., and Hathaway, R. J., "bigVAT: Visual assessment of cluster tendency for large data sets," *Pattern Recognition*, Vol. 38, No. 11, 2005, pp. 1875–1886. URL https://doi.org/10.1016/j.patcog.2005.03.018.

[35] Hathaway, R. J., Bezdek, J. C., and Huband, J. M., "Scalable visual assessment of cluster tendency for large data sets," *Pattern Recognition*, Vol. 39, No. 7, 2006, pp. 1315–1324. URL https://doi.org/10.1016/j.patcog.2006.02.011.

[36] Bezdek, J. C., and Hathaway, R. J., "VAT: A tool for visual assessment of (cluster) tendency," *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, Vol. 3, IEEE, 2002, pp. 2225–2230. URL https://doi.org/10.1109/TFUZZ.2006.889956.

[37] Banerjee, A., and Dave, R., "Validating clusters using the Hopkins statistic," *2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542)*, 2004. URL https://doi.org/10.1109/FUZZY.2004.1375706.

[38] Bezdek, J., Hathaway, R., and Huband, J., "Visual Assessment of Clustering Tendency for Rectangular Dissimilarity Matrices," *IEEE Transactions on Fuzzy Systems, vol. 15, no. 5, pp. 890-903*, 2007. URL https://doi.org/10.1109/TFUZZ.2006.889956.

[39] Statistical Tools For High-Throughput Data Analysis, "Assessing clustering tendency: A vital issue - Unsupervised Machine Learning," , 2008. URL http://www.sthda.com/english/wiki/print.php?id=238#a-single-function-for-hopkins-statistic-and-vat.

[40] Danielsson, P.-E., "Euclidean distance mapping," *Computer Graphics and image processing*, Vol. 14, No. 3, 1980, pp. 227–248. URL https://doi.org/10.1016/0146-664X(80)90054-4.

[41] Wang, L., Zhang, Y., and Feng, J., "On the Euclidean distance of images," *IEEE transactions on pattern analysis and machine intelligence*, Vol. 27, No. 8, 2005, pp. 1334–1339. URL https://doi.org/10.1109/TPAMI.2005.165.

[42] Lange, T., Roth, V., Braun, M. L., and Buhmann, J. M., "Stability-Based Validation of Clustering Solutions," *Neural Computation Vol. 16, 1299-1323*, 2004. URL https://doi.org/10.1162/089976604773717621.

[43] Jun, S., "An Ensemble Method for Validation of Cluster Analysis," *IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 1*, 2011. URL http://www.oalib.com/paper/2647528#.XeQeuy2ZNQI.

[44] Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J., "Understanding of Internal Clustering Validation Measures," *2010 IEEE International Conference on Data Mining, Sydney, NSW, 2010, pp. 911-916.*, 2010. URL https://doi.org/10.1109/ICDM.2010.35.

[45] Statistical tools for high-throughput data analysis (STHDA), "How to choose the appropriate clustering algorithms for your data? - Unsupervised Machine Learning," , 2019. URL http://www.sthda.com/english/wiki/print.php?id=243.

[46] Dunn, J., "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.,vol. 3,pp. 32-57,1973*, 2008. URL https://doi.org/10.1080/01969727308546046.

[47] Rousseeuw, P., "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics Volume 20, Pages 53-65*, 1987. URL https://doi.org/10.1016/0377-0427(87)90125-7.

[48] Ting, K. M., "A Comparative Study of Cost-Sensitive Boosting Algorithms," *In Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, 2000, pp. 983–990. URL https://www.semanticscholar.org/paper/A-Comparative-Study-of-Cost-Sensitive-Boosting-Ting/20212c96cee7b364f2d56aafe53d1d3be5377886.

[49] Freund, Y., and Schapire, R. E., "Schapire R: Experiments with a new boosting algorithm," *In: Thirteenth International Conference on ML*, 1996, pp. 148–156. URL https://www.semanticscholar.org/paper/Experiments-with-a-New-Boosting-Algorithm-Freund-Schapire/68c1bfe375dde46777fe1ac8f3636fb651e3f0f8.

[50] Aslam, J. A., "Improving Algorithms for Boosting," *In Proc. 13th Annu. Conference on Comput. Learning Theory*, Morgan Kaufmann, 2000, pp. 200–207. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.346.

[51] Bauer, E., and Kohavi, R., "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants," *Machine Learning*, 1998, pp. 105–139. URL https://doi.org/10.1023/A:1007515423169.

[52] Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I., and McGregor, G., "Boosted decision trees as an alternative to artificial neural networks for particle identification," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Vol. 543, No. 2, 2005, pp. 577 – 584. doi:https://doi.org/10.1016/j.nima.2004.12.018, URL http://www.sciencedirect.com/science/article/pii/S0168900205000355.

[53] McHugh, M., "Interrater reliability: the kappa statistic," *Biochem Med (Zagreb). ;22(3):276–282*, 2012. URL https://doi.org/10.11613/BM.2012.031.